Collecting Italian spontaneous social media speech: the WAsp2 project

Cenceschi S.[1], Trivilini A.[1], Sbattella L.[2], Tedesco R[2].

[1]The University of Applied Sciences of Southern Switzerland (SUPSI)

[2]ARCSLab, Politecnico di Milano

This work presents the WAsp2 project (WhatsApp SPontaneous SPeech), focused on collecting private social media speech thanks to the WhatsApp application. The main purpose is to collect a wide corpus of spontaneous vocal messages useful in forensic investigations but also in other research topics such as experimental linguistics, speech therapy, artificial intelligence, and ASR systems.

WhatsApp is a *social media* [1] app that allows people to connect with other members, exchanging text messages, photo, audio and video materials. Nowadays, social media constitute a fundamental part of human communications and they are widely used from people of all ages and professions, representing a very large digital pool from which to draw. We think WhatsApp audio messages contributed to introduce a new speech communication style [2][3], characterized by broken conversations and non-consecutive turns, often alternated with text, media, and long temporal breaks. The WAsp2 corpus will be also useful to investigate these communication behaviours and new speakers' expectations compared to the traditional speech.

Many works explore the textual social media language [4][5][6], but far less are those focused on the speech, mostly related to voicemail phone recordings [7][8][9]. A large variety of spontaneous Italian speech corpora have been collected among years [10][11] but none of them focused on social media-style speech. The structure of this work is partially inspired by the *Common Voice project* [12] by Mozilla and the Corpus *Compilation of Private Social Media Messages* [13], but it differs for two main reasons: Wasp2 regards strictly vocal messages and collects spontaneous speech.

Italian native speakers are asked to donate WhatsApp private chats compiling an on-line questionnaire and then sending chat contents by e-mail. We are currently developing the WAsp2 website. A landing page contains a short presentation and the direct link to the authorization to process personal data for research purposes. Once accepted, the donor is redirected to a short questionnaire. We prefer to insert few questions in order to lighten the participant, even if it means losing details. Its purpose is to allow to catalogue recordings according to the following classes.

- Age
- Gender
- Education level
- Geographical origin (where the donor spent most of her/his life)
- Phone model
- If headset has been used ("I don't remember" option is provided)

The selection is not mandatory, but chats without these data will be separate from the others in the final dataset. Once the questionnaire is fulfilled, the screen displays the project e-mail and the summary with WhatsApp instructions to export complete chats or single audio messages through different devices (iPhone/Android/Win phones). This passage can be easily done using the app through the WhatsApp supplied functions. For privacy reason we'll extrapolate only the donor's chat components. Video and images are deleted while audio recordings are divided from the text. Texts are saved for further processing, while audio messages are filtered to keep only the ones containing human speech. ASR techniques are then used to extrapolate transcriptions.

Another webpage will be finally developed in order to validate randomized audio and their corresponding transcriptions. A narrow group of natural language experts will check part of materials; this part is however still being defined.

The "Protocol for the collection of forensics databases" [14] suggests collecting at least two non-contemporaneous recordings of each speaker using different styles, but we prefer to suggest people to send more than 1 chat without further requirements: being them sensitive private data, we prefer to collect as much material as possible, opting for the volunteer's light commitment.

To our knowledge, WAsp2 is the first attempt to ask people to donate their WhatsApp speech so we cannot predict how they will react. An encouraging example is the multilingual large-scale corpus of "What's up, Switzerland?" [4], which collected about 967 textual chats (1 291 022 messages) in four years; it is difficult to make a comparison, however, because people could be more reluctant to donate their voice.

Moreover, we start focusing on the northern Italy and Switzerland varieties reducing the possible number of donors, but being able to rely on a powerful dissemination network. It must be considered that advertising will be more widespread with respect to the previous project, and our intention is to re-launch WAsp2 at fixed periods without a specific deadline.

A first survey, carried out among twenty people, collected only positive responses and returned an interesting suggestion: a crucial aspect is to create a relationship of trust with potential donors. The project will be then spread random, basing the disclosure on public interviews, newsletters to other laboratories or affiliated companies, and exploiting the classic word of mouth among personal contacts. Another possibility is to pay donations, but despite it could be helpful to obtain more chats, we prefer to evaluate the option when the website will be finished.

The forensic research needs to deal with spontaneous speech, but corpora are very limited due to privacy reason. Many datasets are collected among phone services and media (TV, radio) but existing Italian spontaneous speech corpora [15][16] cannot provide enough data to train recognition algorithms. The request for new spontaneous speech recordings is very high, but the only way to make up for its lack is currently to approximate reality with dedicated tasks [17] or using recited/read speech. WAsp2 aims to give a contribute using a new collection methodology. Moreover, WAsp2 environmental recording contexts will be really variable, ensuring different background noise as well as emotional states, gender, age and speakers' speech styles. These characteristics are really useful in the forensic field because they give the opportunity to best approximate real cases. Finally, the WAsp2 corpus will be a possible reference dataset for speaker recognition investigation and Likelihood Ratio calculation, to analyse the influence of quality and audio compression on voice features variations or in investigations related to gender and speech rate.

The corpora will be freely available upon request for research and other non-commercial purposes and gradually updated with new incoming data.

References

[1] Kaplan, A. M. (2015). Social Media, the Digital Revolution, and the Business of Media. International Journal on Media Management, 17(4), 197-199.

[2] Cenceschi, S., & Sbattella, L., Tedesco, R (2018). Verso il riconoscimento automatico della prosodia. STUDI AISV, 433-440.

[3] Nencioni, G. (1983). Di scritto e di parlato. Zanichelli.

[4] Stark, Elisabeth (2016-2018). SNSF project "What's up, Switzerland?" (Sinergia: CRSII1_160714). University of Zurich. www.whatsup-switzerland.ch.

[5] Al-Khawaldeh, N., Bani-Khair, B., Mashaqba, B., & Huneety, A. (2016). A Corpus-Based Discourse Analysis Study of WhatsApp Messenger's Semantic Notifications. International Journal of Applied Linguistics and English Literature, 5(6), 158-165.

[6] Eric N. Forsyth and Craig H. Martell, "Lexical and Discourse Analysis of Online Chat Dialog," Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), pp. 19-26, September 2007.

[7] Chateau, N., Maffiolo, V., & Blouin, C. (2004). Analysis of emotional speech in voice mail messages: The influence of speakers' gender. In Eighth International Conference on Spoken Language Processing.

[8] Koumpis, K., & Renals, S. (2005). Automatic summarization of voicemail messages using lexical and prosodic features. ACM Transactions on Speech and Language Processing (TSLP), 2(1), 1.

[9] Inanoglu, Z., & Caneel, R. (2005, January). Emotive alert: HMM-based emotion detection in voicemail messages. In Proceedings of the 10th international conference on Intelligent user interfaces (pp. 251-253). ACM.

[10] Cresti, E., Moneglia, M., do Nascimento, F. B., Moreno-Sandoval, A., Véronis, J., Martin, P., ... & Blum, C. (2002). The C-ORAL-ROM Project. New methods for spoken language archives in a multilingual romance corpus. In LREC.

[11] 2013 E. Cresti & A. Panunzi, Introduzione ai corpora italiani, Il Mulino, Bologna.

[12] (2017) Mozilla common voice. [Online]. Available: https://voice.mozilla.org/en

[13] Verheijen, L., & Stoop, W. (2016, September). Collecting facebook posts and whatsapp chats. In International Conference on Text, Speech, and Dialogue (pp. 249-258). Springer, Cham.

[14] Morrison, G. S., Rose, P., & Zhang, C. (2012). Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. Australian Journal of Forensic Sciences, 44(2), 155-167.

[15] Emanuela Cresti. Corpus di italiano parlato: Introduzione, volume 1. Accademia della Crusca, 2000.

[16] Federico Albano Leoni. Il corpus clips. presentazione del progetto, 2006.

[17] Rachel Baker and Valerie Hazan. Diapixuk: task materials for the elicitation of multiple spontaneous speech dialogs, 2011.